NTP
National Toxicology Program

Draft OHAT Approach
Part 2
Confidence in the Body of Evidence Through Integrating the Evidence

Andrew Rooney, Ph.D.
Division of the National Toxicology Program,
Office of Health Assessment and Translation (OHAT)

Web-Based Informational Meeting
April 23, 2013 12:00 - 4:00PM EDT

Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments

This Presentation will focus on Steps 5-7

Step 1: Prepare topic

Step 2: Search for and select studies

Step 3: Extract data from studies

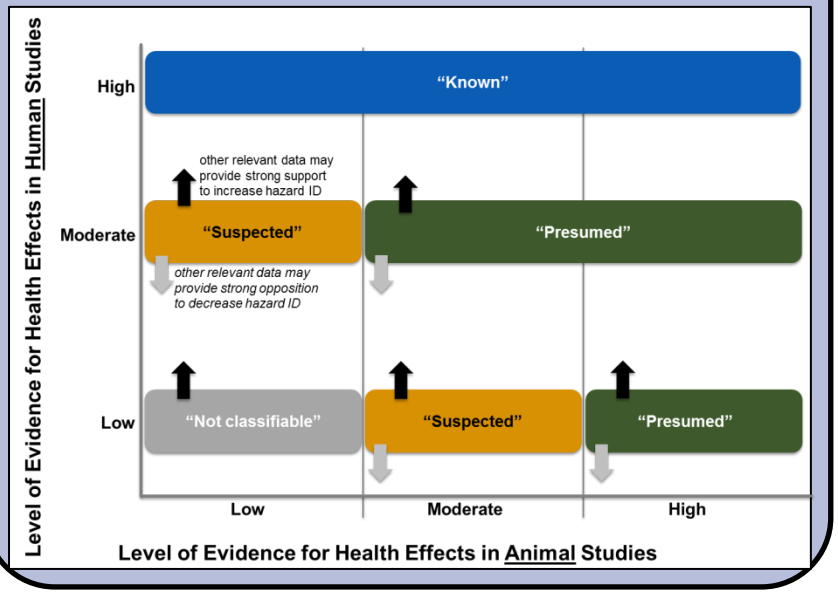
Step 4: Assess individual study quality

Step 5: Rate confidence in body of evidence

Initial Confidence by Key Features of Study Design	Factors Decreasing Confidence	Factors Increasing Confidence	Confidence in the Body of Evidence
High (++++) 4 Features Moderate (+++) 3 Features Low (++) 2 Features Very Low (+) ≤1 Features **Features** • Controlled exposure • Exposure prior to outcome • Individual outcome data • Comparison group used	❖ Risk of Bias ❖ Unexplained Inconsistency ❖ Indirectness ❖ Imprecision ❖ Publication Bias	❖ Large Magnitude of Effect ❖ Dose Response ❖ All Plausible Confounding • Studies report an effect and residual confounding is toward null • Studies report no effect and residual confounding is away from null ❖ Consistency • Across animal models or species • Across dissimilar populations • Across study design types ❖ Other e.g., particularly rare outcomes	High (++++) Moderate (+++) Low (++) Very Low (+)

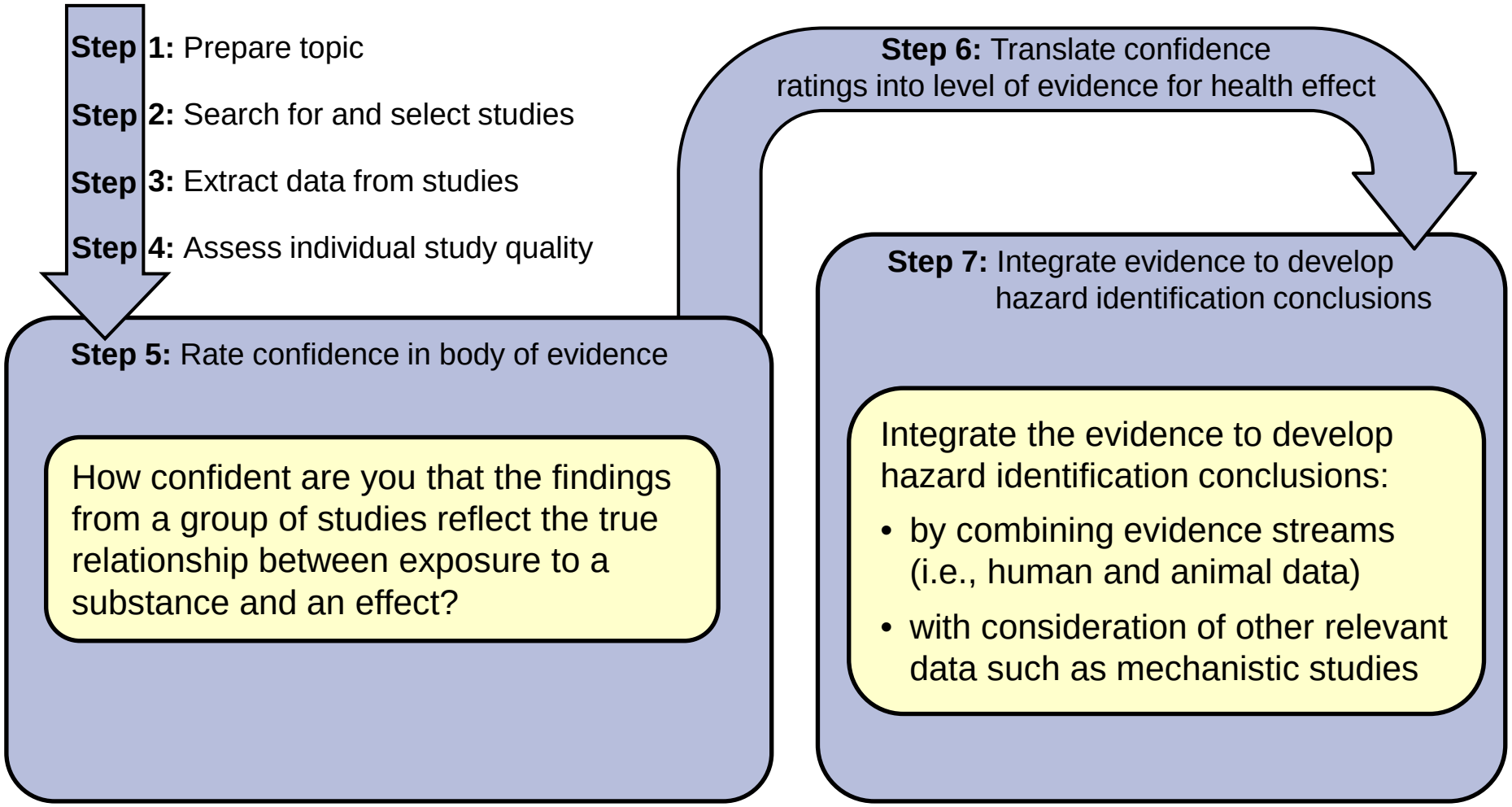
Step 6: Translate confidence ratings into level of evidence for health effect

Step 7: Integrate evidence to develop hazard identification conclusions



Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-Based Health Assessments

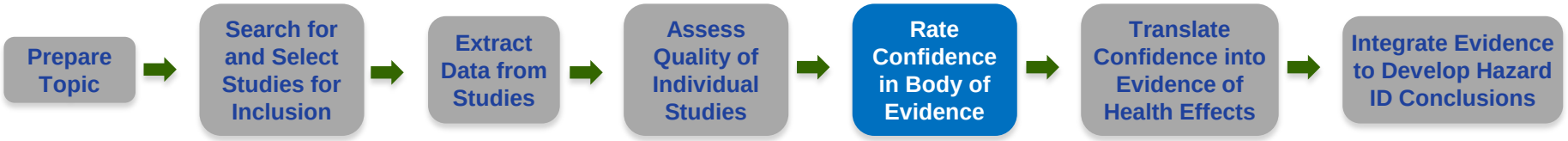
This Presentation will focus on Steps 5-7



Step 5: Rate Confidence in the Body of Evidence

- **Confidence Rating**
  - How confident are you that findings from a group of studies reflect the true relationship between exposure to a substance and an effect?

- **Existing Methods**
  - The GRADE approach is a widely accepted method for rating confidence in a body of evidence
    - No guidance for animal studies
    - No guidance for *in vitro* studies
    - All observational human studies are given the same initial low quality (e.g., case-report = prospective cohort study)



Why GRADE?

- Developed by broad group of international guideline developers in the area of healthcare
- Clear presentation of elements considered for downgrading or upgrading confidence in body of evidence
  - Framework for documenting scientific judgment decisions
  - Elements cover Bradford Hill causality considerations
  - Practitioners engage in ongoing methods development
- Endorsed and used by over 70 organizations
- Consistent with DHHS sister agencies
  - Conceptually similar to AHRQ model
  - Supported by parts of CDC for healthcare recommendations

Step 5: Rate Confidence in the Body of Evidence

- **Confidence Rating** (human and animal data separately)
  - Indicates confidence that findings from the body of evidence reflects the true relationship between exposure to a substance and an effect
  - **Initial Confidence**
    - On an outcome basis
    - Determined by key study design features

Initial Confidence
High
Moderate
Low
Very Low

Key Features
- Controlled exposure
- Exposure prior to outcome
- Individual outcome data
- Comparison group used

Reflect the ability of study design to address confidence that exposure preceded and was associated with outcome

Example:
- **Well conducted experimental studies will have all 4 key features**
- **Therefore "High" initial confidence**

Step 5: Rate Confidence in the Body of Evidence

- **Confidence Rating** (human and animal data separately)
  - Indicates confidence that findings from the body of evidence reflects the true relationship between exposure to a substance and an effect
  - **Initial Confidence**
  - **Factors Decreasing Confidence**

Initial Confidence
High
Moderate
Low
Very Low

Factors Decreasing Confidence
❖ Risk of Bias
❖ Unexplained Inconsistency
❖ Indirectness
❖ Imprecision
❖ Publication Bias

Are there issues that would DECREASE confidence that findings reflect the true relationship between exposure and effects?

Moderate

Example: outcome is indirect measure or "upstream indicator"
- **Decrease confidence from "High" to "Moderate"**

Step 5: Rate Confidence in the Body of Evidence

- **Confidence Rating** (human and animal data separately)
  - Indicates confidence that findings from the body of evidence reflects the true relationship between exposure to a substance and an effect
  - **Initial Confidence**
  - **Factors Decreasing Confidence**
  - **Factors Increasing Confidence**

Initial Confidence	Factors Decreasing Confidence	Factors Increasing Confidence
High	❖ Risk of Bias	❖ Large Magnitude of Effect
Moderate	❖ Unexplained Inconsistency	❖ Dose Response
Low	❖ Indirectness	❖ All Plausible Confounding
Very Low	❖ Imprecision	❖ Consistency
	❖ Publication Bias	❖ Other

Are there issues that INCREASE confidence that findings reflect the true relationship between exposure and effects?

→ Moderate

<u>Example:</u> no issues
- **No increase in confidence**

Step 5 Schematic: Adaptations to Address Breadth of Data Relevant for Environmental Health Questions

Initial confidence set by study design features in OHAT Approach (stratifies observational studies)

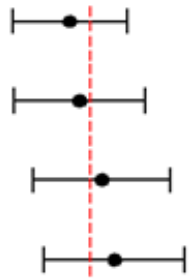
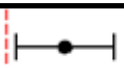
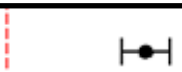
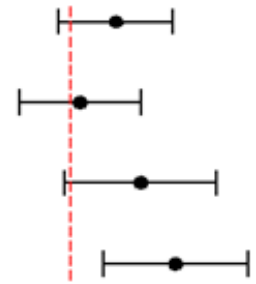
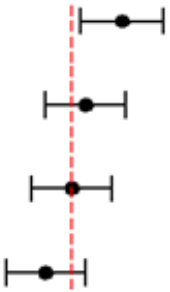
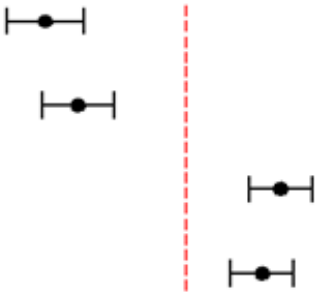
Initial Confidence by Key Features of Study Design	Factors Decreasing Confidence	Factors Increasing Confidence	Confidence in the Body of Evidence
High (++++) **4 Features**	❖ **Risk of Bias** ❖ **Unexplained Inconsistency** ❖ **Indirectness** ❖ **Imprecision** ❖ **Publication Bias**	❖ **Large Magnitude of Effect** ❖ **Dose Response** ❖ **All Plausible Confounding** • Studies report an effect and residual confounding is toward null • Studies report no effect and residual confounding is away from null ❖ **Consistency** • Across animal models or species • Across dissimilar populations • Across study design types ❖ **Other** e.g., particularly rare outcomes	High (++++)
Moderate (+++) **3 Features**			Moderate (+++)
Low (++) **2 Features**			Low (++)
Very Low (+) **≤1 Features**			Very Low (+)

Features
- Controlled exposure
- Exposure prior to outcome
- Individual outcome data
- Comparison group used

OHAT added consistency across breadth of data

Example Guidance in Protocols:
When to Downgrade for Unexplained Inconsistency

Table 15. Factors to consider when considering consistency of results

No downgrade		
– Point estimates similar – Confidence intervals overlap – Statistical heterogeneity is non-significant – I^2 of ≤50%		
Example A		
		
χ^2 p-level = 0.767; I^2 = <<1%; τ^2 = <<1	χ^2 p-level = 0.017; I^2 = 71%; τ^2 = 0.044	χ^2 p-level = <0.001; I^2 = 98%; τ^2 = 1.022
Example B	Example B	Example B
		
χ^2 p-level = 0.241; I^2 = 29%; τ^2 = 0.046	χ^2 p-level = 0.068; I^2 = 58%; τ^2 = 0.025	χ^2 p-level = <0.001; I^2 = 98%; τ^2 = 0.774

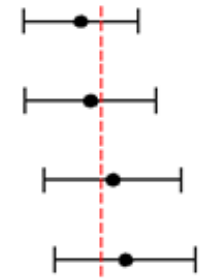
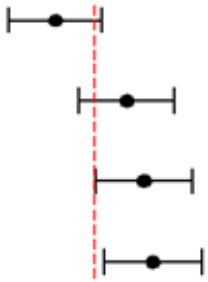
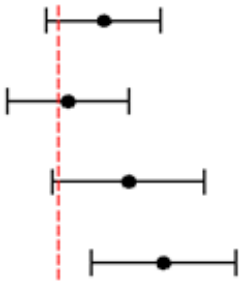
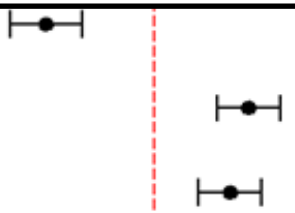
Factors to Consider for No Downgrade

- Point estimates similar
- Confidence intervals overlap
- Statistical heterogeneity non-significant (p≥0.1)
- I^2 of ≤50%
- Example figures

*protocol also includes guidance on when we might conduct a quantitative data synthesis

Example Guidance in Protocols:
When to Downgrade for Unexplained Inconsistency

Table 15. Factors to consider when considering consistency of results

No downgrade	One level downgrade (serious)	
– Point estimates similar – Confidence intervals overlap – Statistical heterogeneity is non-significant – I^2 of ≤50%	– Point estimates vary – Confidence intervals show minimal ove – Statistical heterogeneity has low p-valu – I^2 of >50% to 75%	
Example A  χ^2 p-level = 0.767; I^2 = <<1%; τ^2 = <<1	Example A  χ^2 p-level = 0.017; I^2 = 71%; τ^2 = 0.044	
Example B  χ^2 p-level = 0.241; I^2 = 29%; τ^2 = 0.046	Example B  χ^2 p-level = 0.068; I^2 = 58%; τ^2 = 0.025	 χ^2 p-level = <0.001; I^2 = 98%; τ^2 = 0.774

Factors to Consider to Downgrade 1 Level

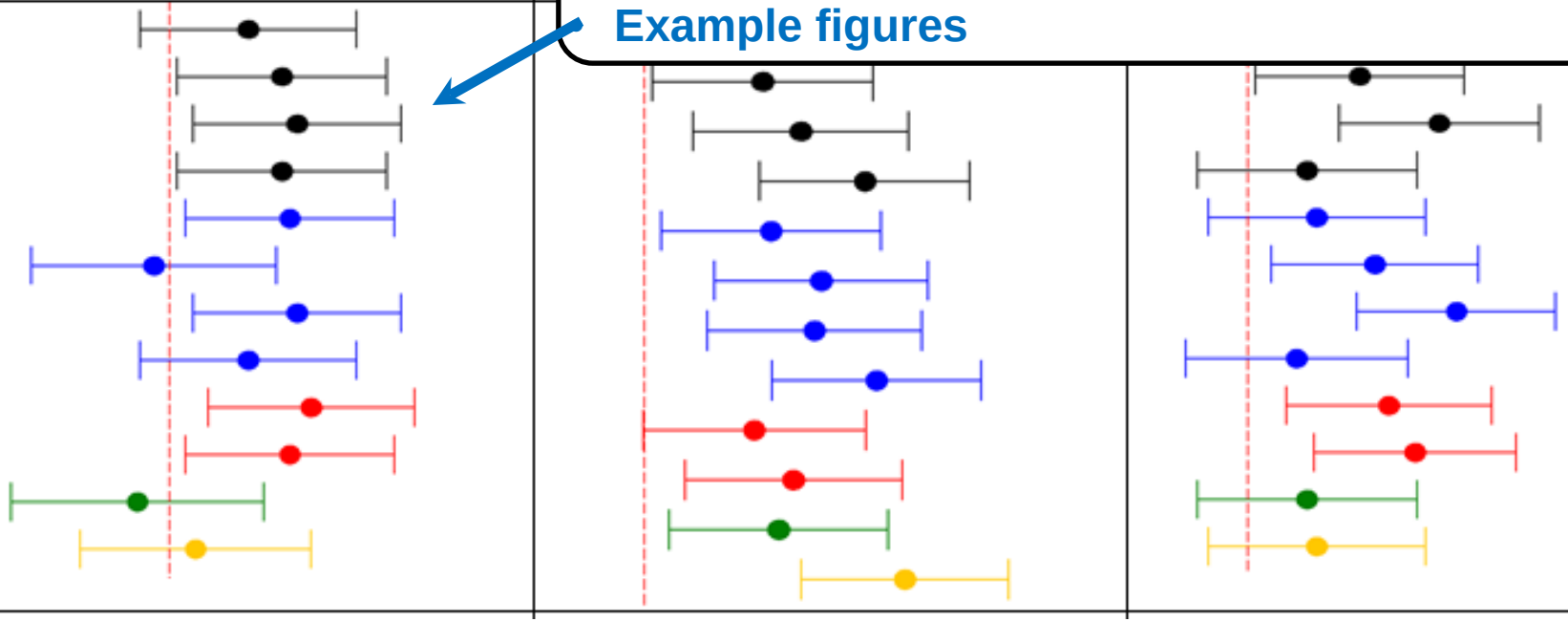
- Point estimates vary
- Confidence intervals show minimal overlap
- Statistical heterogeneity has low p-value (p<0.1)
- I^2 of >50% to 75%
- Example figures

*protocol also includes guidance on when we might conduct a quantitative data synthesis

Example Guidance in Protocols:
When to Upgrade for Dose Response Gradient

Table 19. Conceptual examples of upgrade decisions for evidence of dose response gradient

no upgrade		
Example A, findings sorted by study and then dose or exposure level (low to high)		nd gh)



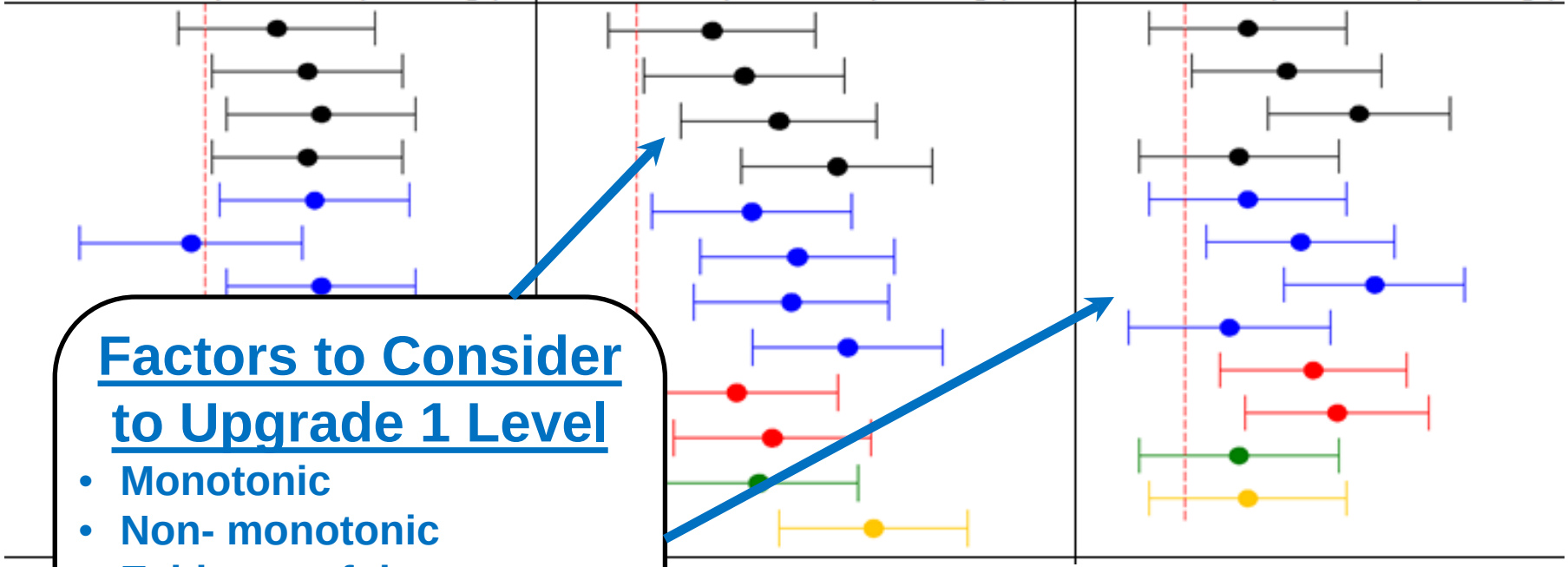
Factors to Consider for No Upgrade

- No evidence of dose response
- Example figures

Example Guidance in Protocols:
When to Upgrade for Dose Response Gradient

Table 19. Conceptual examples of upgrade decisions for evidence of dose response gradient

no upgrade	upgrade +1 (monotonic)	upgrade +1 (non-monotonic)[1]
Example A, findings sorted by study and then dose or exposure level (low to high)	Example B, findings sorted by study and then dose or exposure level (low to high)	Example C, findings sorted by study and then dose or exposure level (low to high)



Factors to Consider to Upgrade 1 Level

- Monotonic
- Non- monotonic
- Evidence of dose response within a study
- Evidence of dose response across studies

Reaching Final Confidence Conclusions on Human and Animal Studies

- Conclusions are based on the evidence with the highest confidence rating when considering across study designs and multiple outcomes

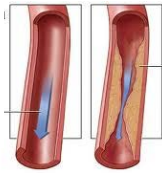
- **Across biologically-related outcomes**
  - **First**: rate confidence in individual outcomes
  - **Then**: re-evaluate confidence conclusion for combined outcomes
  - The overall confidence conclusion for a combined outcome can differ from (e.g., be higher than) the individual outcome ratings

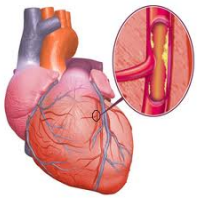
Example:

Blood Pressure

Cardiovascular disease

Cardiovascular mortality





- **Note:** If body of evidence has “Very Low” confidence, it is not used to develop hazard ID conclusions in steps 6 and 7

Confidence in Other Relevant Studies: Assessment of Biological Plausibility

Factors considered when evaluating the support for biological plausibility provided by *in vitro*, cellular, genomic, or mode of action data

Strong Support[1] **Weak Support**

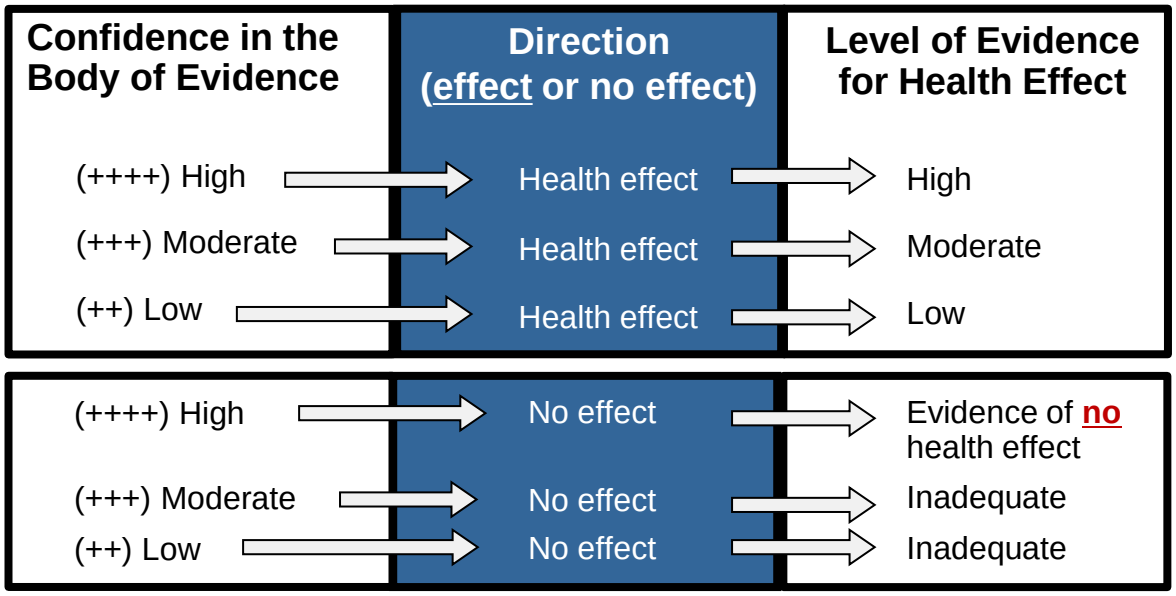
- ***Relevance of biological process or pathway to human health***
- ***Consistency***
- ***Relevance of concentration***
- ***Potency***
- ***Dose response***
- ***Publication bias***

Factors considered parallel elements used to evaluate confidence in the other data streams

A conclusion of "strong" support for biological plausibility requires that most elements are met

Step 6: Translate Confidence Ratings into Level of Evidence for Health Effects

- **Level of evidence for health effects** conclusions reflect
  - The overall confidence in the association between exposure to a substance and a given outcome, and
  - The direction of the effect (toxicity or no toxicity)



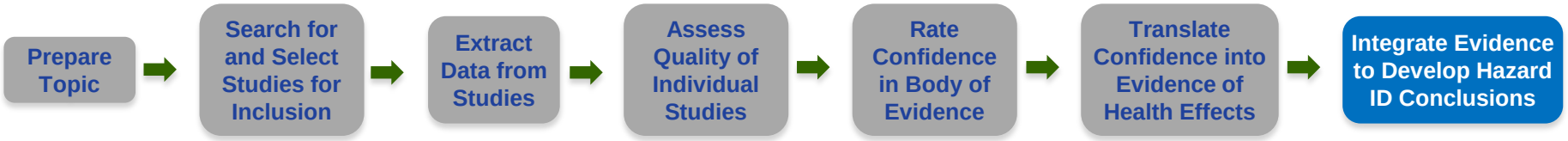
Note: descriptors are applied separately to human and experimental animal evidence

Prepare Topic → Search for and Select Studies for Inclusion → Extract Data from Studies → Assess Quality of Individual Studies → Rate Confidence in Body of Evidence → **Translate Confidence into Evidence of Health Effects** → Integrate Evidence to Develop Hazard ID Conclusions

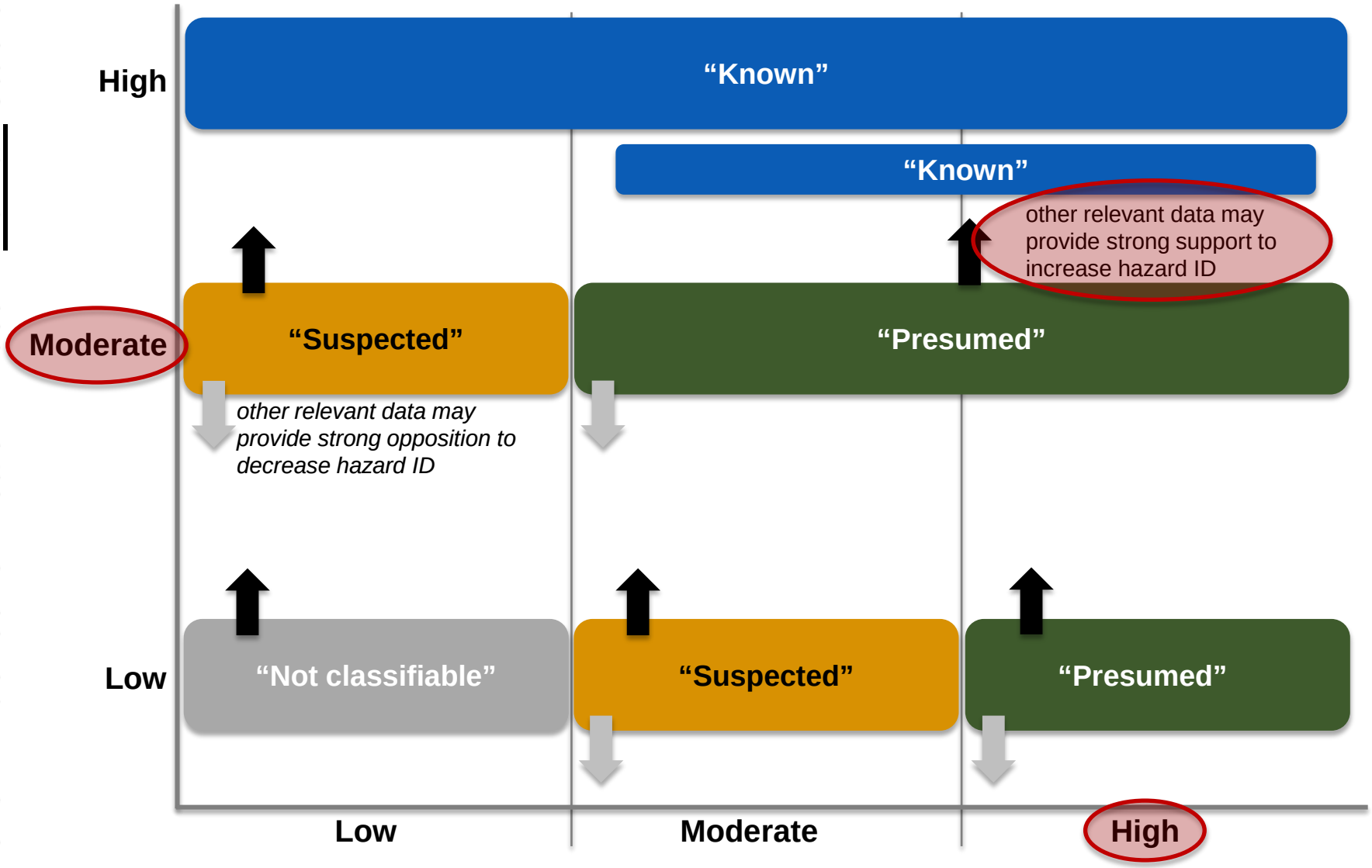
Step 7: Integrate Evidence to Develop Hazard Identification Conclusions

- **Integrate evidence** by combining evidence streams to reach one of four overall hazard identification conclusions
  - **Known** to be a hazard to humans
  - **Presumed** to be a hazard to humans
  - **Suspected** to be a hazard to humans
  - **Not classifiable** to be a hazard to humans

- Two part process for integrating the evidence
  - Consider human evidence and animal evidence together
  - Consider impact of other relevant data
    - e.g., mechanistic, *in vitro*, or upstream indicator data



Integrate Evidence to Develop Hazard ID Conclusions



Assessment of Biological Plausibility Provided by Other Relevant Studies: PFOA/PFOS and Immunotoxicity

- **Consider upgrading the hazard ID**
  If other relevant data provide strong support for biological plausibility of the relationship between exposure and the health effect
  - To provide support, the mechanistic or *in vitro* data must support biological plausibility of observed immune outcomes from human epidemiology or *in vivo* animal studies
  - It is also envisioned that strong evidence for a relevant biological process from mechanistic or *in vitro* data could result in a conclusion of "suspected" in the absence of human epidemiology or *in vivo* animal data

Assessment of Biological Plausibility Provided by Other Relevant Studies: PFOA/PFOS and Immunotoxicity

Factors considered when evaluating the support for biological plausibility provided by *in vitro*, cellular, genomic, or mode of action data

More detail and examples provided in the protocol

Strong Support[1]

- ***Relevance of biological process or pathway to human health***
  generally accepted as relevant (e.g., myelotoxicity or bone marrow toxicity)
- ***Consistency***
  consistency across multiple studies (preferably in more than 2 in different model systems for the same biological pathway)
- ***Relevance of concentration***
  physiologically relevant or "low" concentration effects (e.g., mean of 3-5ng/ml PFOA and 9–30 ng/ml PFOS in the US population 1999-2010 (CDC 2012) range of 17-5100 ng/ml PFOA and 37-3490 ng/ml PFOS in occupationally exposed adults)
- ***Potency***
  magnitude of response
- ***Dose response***
  displays expected dose
- ***Publication bias***
  undetected

Consistency still applies in absence of *in vivo* data, analogous to other data streams

Consistency

- Within context of observed *in vivo* immune outcomes
  - IgE supports sensitization
  - IgE does not support NK
- Stronger if data provide information on multiple steps along the relevant biological pathway
- Also applies to repeatability within the same assay across studies

Causality Considerations in draft OHAT Approach

Hill Considerations	Consideration in the OHAT Approach
Strength	• upgrading the confidence in the body of evidence for ***large magnitude of effect*** • downgrading confidence for **imprecision**

Consistency	• upgrading confidence in the body of evidence for
	  • ***consistency across study types***,
	  • ***consistency across dissimilar populations***
	  • ***consistency across animal species or models***
	• integrating the body of evidence among human, animal, and other relevant data
	• downgrading confidence in the body of evidence for ***unexplained inconsistency***

Temporality	• the ***initial confidence ratings*** by study design, for example experimental studies are rated "High" because of the increased confidence that exposure preceded outcome
Biological gradient	• upgrading the confidence in the body of evidence for a ***dose-response*** relationship
Biological plausibility	• in examining non monotonic **dose-response** relationships • in developing confidence conclusions across biologically related outcomes • other relevant data that inform plausibility are considered in integrating the body of evidence • downgrading the confidence in the body of evidence for **indirectness**
Experimental evidence	• the **initial confidence ratings** by study design • downgrading for **risk of bias**

GRADE consideration discussed in Schunemann et al., 2011: The GRADE approach and Bradford Hill's criteria for causation. JECH 65: 392.

Next Steps

- Framework is currently available for public comment
  - Released publically February 25, 2013
  - For more files and details see http://ntp.niehs.nih.gov/go/38673
  - **Public comment period ends June 11, 2013**
- Two case studies to assess and refine methods
  - Protocols illustrate the application of this framework
    - BPA exposure and obesity
    - PFOA or PFOS exposure and immunotoxicity
    - Released publically April 9, 2013
- Careful consideration of comments from public and at NTP Board of Scientific Councilors Meeting June 25, 2013
- Release updated guidance
  - Expect to be updated periodically, e.g., new best practices

Acknowledgements

- **Office of Health Assessment and Translation**
  - Abee Boyles
  - Kembra Howdeshell
  - Andrew Rooney, Deputy Director
  - Michael Shelby
  - Kyla Taylor
  - Kristina Thayer, Director
  - Vickie Walker
- **Office of Liaison, Policy and Review**
  - Mary Wolfe, Director
  - Lori White
- **Approach Technical Advisors and Experts**
  - **Lisa Bero,** Director, San Francisco Branch, United States Cochrane Center at UC San Francisco
  - **Gordon Guyatt,** Co-chair, GRADE Working Group, McMaster University
  - **Malcolm Macleod,** CAMARADES Centre, University of Edinburgh
  - **Karen Robinson,** Co-Director, Evidence-Based Practice Center, The Johns Hopkins Bloomberg School of Public Health
  - **Holger Schünemann**, Co-chair, GRADE Working Group, McMaster University
  - **Tracey Woodruff,** Director, Program on Reproductive Health and the Environment, UCSF
- **NTP BSC Working Group**
  - **Lynn Goldman, Chair,** Dean, School of Public Health and Health Services, George Washington University, Washington, DC
  - **Reeder Sams, Vice-chair,** Acting Deputy Director, National Center for Environmental Assessment/RTP Division, USEPA
  - **Lisa Bero,** Director, San Francisco Branch, United States Cochrane Center at UC San Francisco
  - **Edward Carney,** Senior Science Leader, Mammalian Toxicology, Dow Chemical Company
  - **David Dorman,** Professor, North Carolina State University
  - **Elaine Faustman**, Director, Institute for Risk Analysis and Risk Communication, University of Washington
  - **Dale Hattis**, Research Professor, George Perkins Marsh Institute, Clark University
  - **Malcolm Macleod,** CAMARADES Centre, University of Edinburgh
  - **Tracey Woodruff,** Director, Program on Reproductive Health and the Environment, UCSF
  - **Lauren Zeise,** Chief, Reproductive and Cancer Hazard Assessment Branch, OEHHA, California EPA
- **Protocol Technical Advisors**

1: Prepare Topic

2: Search for and Select Studies for Inclusion

3: Extract Data from Studies

4: Assess Quality of Individual Studies

- ++ Definitely Low risk of bias
- + Probably Low risk of bias
- – Probably High risk of bias
- –– Definitely High risk of bias

5: Rate Confidence in the Body of Evidence

Initial Confidence by Key Features of Study Design	Factors Decreasing Confidence	Factors Increasing Confidence	Confidence in the Body of Evidence
High (++++) 4 Features	❖ Risk of Bias ❖ Unexplained Inconsistency ❖ Indirectness ❖ Imprecision ❖ Publication Bias	❖ Large Magnitude of Effect ❖ Dose Response ❖ All Plausible Confounding • Studies report an effect and residual confounding is toward null • Studies report no effect and residual confounding is away from null ❖ Consistency • Across animal models or species • Across dissimilar populations • Across study design types ❖ Other e.g., particularly rare outcomes	High (++++)
Moderate (+++) 3 Features			Moderate (+++)
Low (++) 2 Features			Low (++)
Very Low (+) ≤1 Features			Very Low (+)

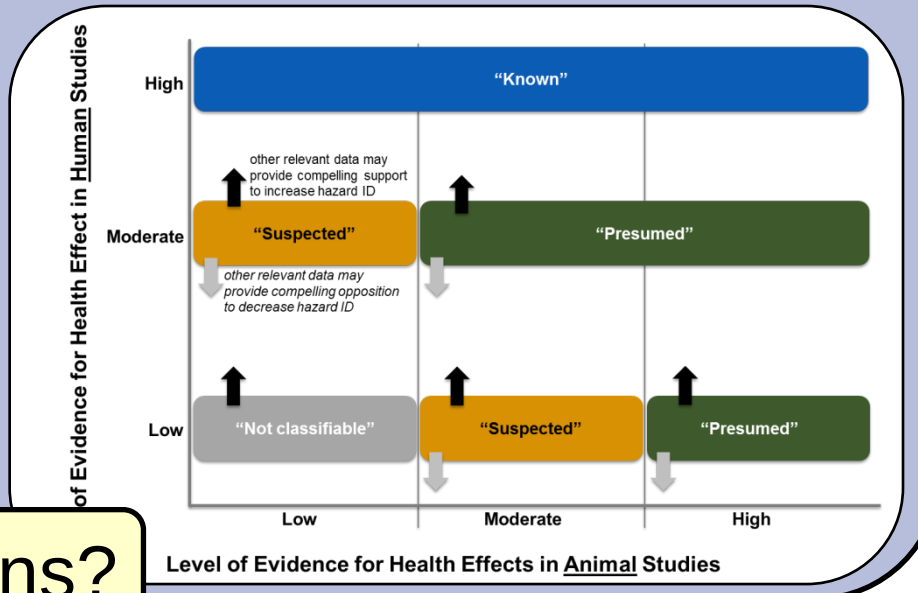
Features
- Controlled exposure
- Exposure prior to outcome
- Individual outcome data
- Comparison group used

Questions?

6: Translate Confidence Ratings into Evidence of Health Effects

Confidence in the Body of Evidence	Direction (effect or no effect)	Level of Evidence for Health Effect
(+++) High	Health effect	High
(++) Moderate	Health effect	Moderate
(+) Low	Health effect	Low
(+++) High	No effect	Evidence of no health effect
(++) Moderate	No effect	Inadequate
(+) Low	No effect	Inadequate

7: Integrate Evidence to Develop Hazard ID Conclusions



Extra Slides

Example Guidance in Protocols:
When to Downgrade for Indirectness

Table 15. Guidance for downgrading human studies for directness				
Health outcomes		**Exposure scenario**	**Time between exposure and outcome assessment**	**Overall downgrade**
primary	0	0	0	0
secondary	-1	0	0	-1
0 = no downgrade, -1 = one downgrade, -2 two downgrade				

Example Guidance in Protocols:
When to Downgrade for Indirectness
PFOA / PFOS Exposure and Immunotoxicity

Table 16. Guidance for downgrading animal studies for directness

Animal model		Health outcomes		Route of administration		Time between treatment and assessment	Overall downgrade
Mammalian	0	primary	0	oral, sc injection, dermal, inhalation	0	0	0
				intraperitoneal injection	-1	0	-1
		secondary	-1	oral, injection, dermal, inhalation	0	0	-1
				Intraperitoneal (ip) injection	-1	0	-2
Non-mammalian vertebrates	-1	primary	0	oral, sc injection, dermal, inhalation	0	0	-1
				ip, water for aquatic species	-1	0	-2
		secondary	-1				
Invertebrates	-2	primary	0				
		secondary	-1				

0 = no downgrade, -1 = one downgrade, -2 two downgrade
sc = subcutaneous, ip = intraperitoneal

- Route of administration

Downgrade for Indirectness

- Model (mammal=0, vertebrate -1, invertebrate -2)
- Health outcome (primary = 0, secondary -1)

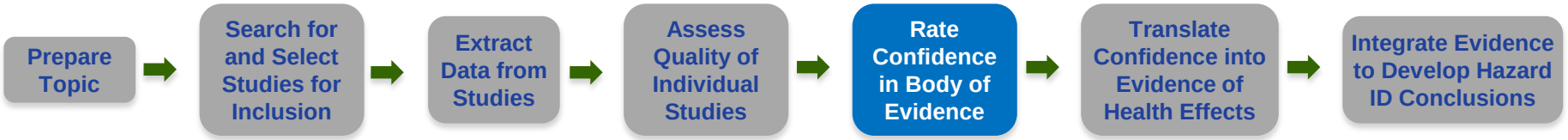
Key Study Design Features for Initial Confidence

1. **Exposure to the substance is controlled**
   - Experimental studies can largely eliminate confounding by randomizing allocation of exposure

2. **Exposure assessment represents exposures occurring prior to the development of the outcome**
   - Supports causal pathway and if present, it is unlikely that association is the result of reverse causation

3. **Outcome is assessed on the individual level** (i.e., not population aggregate data)
   - Without individual-level information on outcomes, a study cannot control for additional confounding variables ("ecologic fallacy")

4. **Comparison group is used within the study** (e.g., not case reports)



Example Details Included in Summary Tables

Table 6 from PFOA/PFOS Exposure and Immunotoxicity Protocol

Reference, Study Design & Population	Health Outcome	Exposure	Statistical Analysis	Results
(Carwile and Michels 2011) **Study Design:** cross-sectional Adults who participated in the 2003/04 and 2005/06 National Health and Nutrition Examination Survey (NHANES) and a spot urine sample analysed for BPA. **N:** 2747 **Location:** US, NHANES national survey **Sex (% male):** ♂♀(49.6%) **Sampling time frame:** 2003-2006 **Age:** 18-74 years **Exclusions:** pregnant women, participants with missing urinary BPA, creatine, BMI, or covariate data **Funding Source:** NIH National Research Service (NRSA) **Author conflict of interest:** not reported	**Diagnostic and prevalence in total cohort:** **obesity:** BMI ≥ 30 (n=932, 34.3%) **overweight:** 25 ≤ BMI < 30 (n=864, 31.8%) **elevated waist circumference (WC):** >102 cm in ♂ or ≥ 88 cm in ♀ (n=1330, 50%) *BMI = body mass index (kg/m^2)	**Exposure assessment:** urine (µg/g creatinine or ng/ml and creatinine as adjustment variable) measured by online SPE-HPLC-MS/MS (Ye 2005) **Exposure levels:** 2.05 µg/g creatinine (geometric mean), 1.18-3.33 (25-75th percentile) Q1: ≤1.1 ng/ml Q2: 1.2-2.3 ng/ml Q3: 2.4-4.6 ng/ml Q4: ≥4.7 ng/ml	**obesity & overweight:** polytomous regression **elevated WC:** logistic regression **Adjustment factors:** sex, age, race, urinary creatinine, education, smoking **Statistical power:** "appears to be adequately powered" based on ability to detect an OR of 1.5 with 80% power using Q1 prevalence of 40.4% obesity, 44.4% overweight, and 46% elevated WC	adjOR (95% CI) ***obesity*** Q2 vs Q1: 1.85 (1.22,2.79) Q3 vs Q1: 1.60 (1.01,2.44) Q4 vs Q1: 1.76 (1.06,2.94) ***overweight*** Q2 vs Q1: 1.66 (1.21,2.27) Q3 vs Q1: 1.26 (0.85,1.87) Q4 vs Q1: 1.31 (0.80,2.14) ***elevated WC*** Q2 vs Q1: 1.62 (1.11,2.36) Q3 vs Q1: 1.39 (1.02,1.90) Q4 vs Q1: 1.58 (1.03,2.42)

statistical power as "appears to be adequately powered" (sample size met), somewhat underpowered (sample size is 75% to <100% of recommended), "underpowered" (<75% required), or "severely underpowered (sample size is <50% required)

RISK OF BIAS ASSESSMENT

Risk of bias response options for individual items: should we delete domains from this table?

Bias Domain	Criterion		Re
Selection	Was administered dose or exposure level adequately randomized?	n/a	not applicable
	Was allocation to study groups adequately concealed?	n/a	not applicable
	Were the comparison groups appropriate?	++	yes, based on qu
Confounding	Does the study design or analysis account for important confounding and modifying variables?	++	yes (sex, age, race adjustment for nut
	Did researchers adjust or control for other exposures that are anticipated to bias results?	+	no, but not consid tudies
Performance	Were experimental conditions identical across study groups?		
	Did deviations from the study protocol impact the results?		
	Were the research personnel and human subjects blinded to the study group during the study?		
Attrition	Were outcome data incom		or any analysis)
Detection	Were the outcome assesso		sessment
	Were confounding variable		
	Can we be co	++	yes, NHANES methods are considered "gold standard" for urinary BPA
	Can we be co	++	yes, used standard diagnostic criteria
Selective Reporting	Were all mea	++	yes, primary outcomes discussed in methods were presented results section with adequate level of detail for data extraction
Other	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?	++	none identified

RISK OF BIAS — 1st Tier for risk of bias

Risk of bias response options for individual items:	
++	definitely low risk of bias
+	probably low risk of bias
-	probably high risk of bias
--	definitely high risk of bias
n/a	not applicable

Results

Analysis

Exposure

Health Outcome

Reference, Study Design and Population

Risk of Bias

Example Risk of Bias Details in Summary Table

Table 6 from PFOA/PFOS R

Reference, Study Design & Population
(Carwile and Michels 2011) **Study Design:** cross-sectional Adults who participated in the 2003/04 and 2 National Health and Nutrition Examination Su (NHANES) and a spot urine sample analysed f **N:** 2747 **Location:** US, NHANES national survey **Sex (% male):** ♂♀(49.6%) **Sampling time frame:** 2003-2006 **Age:** 18-74 years **Exclusions:** pregnant women, participants wi missing urinary BPA, creatine, BMI, or covaria **Funding Source:** NIH National Research Servi **Author conflict of interest:** not reported
statistical power as "appears to be adequately required), or "severely underpowered (sample size

Risk of Bias

- Rating/answer to applicable questions
- Answers justified with text from study
- Hypothetical example on confounding:

"yes (sex, age, race urinary creatinine, education, smoking), but no adjustment for nutritional quality"

RISK OF BIAS ASSESSMENT			
Risk of bias response options for individual items: should we delete domains from this table?			
Bias Domain	**Criterion**	**Response**	
Selection	Was administered dose or exposure level adequately randomized?	n/a	not applicable
	Was allocation to study groups adequately concealed?	n/a	not applicable
	Were the comparison groups appropriate?	++	yes, based on quartiles of exposure
Confounding	Does the study design or analysis account for important confounding and modifying variables?	++	yes (sex, age, race, urinary creatinine, education, smoking), but no adjustment for nutritional quality, e.g., soda consumption
	Did researchers adjust or control for other exposures that are anticipated to bias results?	+	no, but not considered to present risk of bias in general population studies
Performance	Were experimental conditions identical across study groups?	n/a	not applicable
	Did deviations from the study protocol impact the results?	+	no deviations reported
	Were the research personnel and human subjects blinded to the study group during the study?	n/a	not applicable
Attrition	Were outcome data incomplete due to attrition or exclusion from analysis?	+	not considered a risk of bias, excluded observations (≤ 87 for any analysis) based on missing BMI or covariate data
Detection	Were the outcome assessors blinded to study group or exposure level?	++	yes, BPA levels not known at time of outcome assessment
	Were confounding variables assessed consistently across groups using valid and reliable measures?	++	yes, used standard NHANES methods
	Can we be confident in the exposure characterization?	++	yes, NHANES methods are considered "gold standard" for urinary BPA
	Can we be confident in the outcome assessment?	++	yes, used standard diagnostic criteria
Selective Reporting	Were all measured outcomes reported?	++	yes, primary outcomes discussed in methods were presented results section with adequate level of detail for data extraction
Other	Were there any other potential threats to internal validity (e.g., inappropriate statistical methods)?	++	none identified
RISK OF BIAS		1st Tier for risk of bias	

Risk of bias response options for individual items:	
++	definitely low risk of bias
+	probably low risk of bias
-	probably high risk of bias
--	definitely high risk of bias
n/a	not applicable